Proc. 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2010), Desenzano del Garda, Italy, April 2010.

# LOCAL INVARIANT FEATURE TRACKS FOR HIGH-LEVEL VIDEO FEATURE EXTRACTION

*Vasileios Mezaris, Anastasios Dimou, Ioannis Kompatsiaris*

Informatics and Telematics Institute / Centre for Research and Technology Hellas
6th Km Charilaou-Thermi Road, Thermi 57001, Greece
{bmezaris, dimou, ikom}@iti.gr

## ABSTRACT

This paper builds upon previous work on local interest point detection and description to propose the extraction and representation of novel Local Invariant Feature Tracks (LIFT). These features compactly capture not only the spatial attributes of 2D local regions, as in SIFT and related techniques, but also their long-term trajectories in time. This and other desirable properties of LIFT allow the generation of Bags-of-Spatiotemporal-Words models that facilitate capturing the dynamics of video content, which is necessary for detecting high-level video features that by definition have a strong temporal dimension. Preliminary experimental evaluation and comparison of the proposed approach reveals promising results.

## 1. INTRODUCTION

The development of algorithms for the automatic understanding of the semantics of multimedia and in particular of video content, and the semantic indexing by means of high-level features corresponding to semantic classes (objects, events) is currently one of the major challenges in multimedia research. This is motivated by the ever-increasing pace at which video content is generated, rendering any annotation scheme that requires human labor unrealistically expensive and unpractical for use on anything but a very restricted subset of the generated content that may be of unusually high value or importance (e.g. cinema productions, medical content).

Research efforts towards the goal of high-level video feature extraction have followed in the last decade or so several different directions that have the potential to contribute to this goal, ranging from segmentation and key-frame extraction to video content representation using global shot or image features, local interest point detection and description [1], creation of visual lexicons for video representation (Bag-of-Words [2]), machine learning for associating low-level and high-level features, etc. Typically, techniques belonging to several of the aforementioned categories need to be carefully combined for extracting high-level video features.

This work focuses on video content representation, and in particular builds upon previous work on local interest point detection and description to propose the extraction and representation of Local Invariant Feature Tracks (LIFT). These features compactly describe the appearance and the long-term motion of local regions and are invariant, among others, to camera motion, in contrast to both 2D interest point descriptors and their known extensions to spatio-temporal interest points. The proposed feature tracks are shown to be suitable for the generation of a Bag-of-Spatiotemporal-Words model that facilitates capturing the dynamics of video content, allowing the more reliable detection of high-level features that have a strong temporal dimension (e.g. "people-dancing").

The rest of the paper is organized as follows: in section 2, previous work on local interest point detection and description is discussed; in section 3, the proposed LIFT representation is presented; preliminary experimental results are reported in section 4 and finally conclusions are drawn in section 5.

This work was supported by the European Commission under contracts FP6-045547 VIDI-Video and FP7-248984 GLOCAL.

## 2. RELATED WORK

Several approaches to scale-invariant interest point detection and description in still images have been proposed and are widely used in still image understanding tasks (image classification, object detection, etc.) as well as other applications. SIFT [1] is probably the most widely adopted method; SIFT-based descriptors are shown in [3] to outperform several previously proposed techniques for local region description. More recent work on this topic includes SURF [4], which focuses mostly on speeding-up the interest point detection and description process, and [5], which examines the introduction of color information to the original grey-value SIFT. For the application of high-level feature extraction in generic image collections, the above descriptors are typically used to build a Bag-of-Words (BoW) model, which involves the definition of a "vocabulary" of visual words (typically by clustering the interest point descriptors coming from a large number of images and selecting the resulting centroids as words) and the subsequent representation of each image as the histogram of the visual words (i.e. corresponding interest points) found in it.

Large scale video analysis for the purpose of high-level feature extraction, using local invariant features, is in most cases performed at the key-frame level [6]. Thus, the video analysis task reduces to still image analysis. This has obvious advantages in terms of computational complexity, but on the other hand completely disregards the temporal dimension of video and the wealth of information that is embodied in the evolution of the video frames along time. The temporal evolution of the video signal, i.e. motion, is generally considered to convey very important information in video, being a key element of several video understanding and manipulation tasks. Long-term region trajectories in particular, rather than the motion at the frame level, was shown in many works to be very useful in video segmentation, indexing and retrieval (e.g. [7]). Similarly to other analysis tasks, the use of video data in excess of one single key-frame (e.g. using multiple key-frames per shot [8], or treating all frames as key-frames and also considering their temporal succession [9]) for high-level feature extraction has been shown to lead to improved results.

In order to introduce temporal information in the interest-point-based representation of video shots, the use of spatio-temporal (as opposed to spatial-only) interest point detectors has been proposed [10]. Spatio-temporal interest points are defined as locations in the video where intensity values present significant variations both in space and in time. In [11] and other works, such points are used for human action categorization, since the abrupt changes in motion that trigger the detection of spatio-temporal interest points can be useful in discriminating between different classes of human activity (walking, jumping, etc.). However, spatio-temporal interest points define 3D volumes in the video data that typically neither account for possible camera motion nor capture long-term local region trajectories. To alleviate these drawbacks, the tracking of spatial interest points across successive frames has been proposed for applications such as object tracking [12]. For retrieving similar shots or objects within a video, the tracking of SIFT features and the clustering of the resulting tracks of a shot into clusters corresponding to objects is proposed in [13]. In [14], interest points are tracked and the motion information alone (i.e. the trajectories) are used for describing the motion patterns that are present in a sequence of frames, for the purpose of human action recognition and event-based video retrieval.

## 3. LOCAL INVARIANT FEATURE TRACKS

### 3.1. Feature Track extraction

Let $S$ be a shot comprising $T$ frames, $S = \{I_t\}_{t=0}^{T-1}$, coming from the temporal sub-sampling of the original video shot $S^0 = \{I_\tau\}_{\tau=0}^{T^0-1}$ by a factor of $a$; $T = \lceil T^0/a \rceil$.

Application of any of the available combinations of interest point detectors and descriptors (e.g. [1, 4, 5]) on a frame $I_t$ of $S$ results in the extraction of a set of interest point descriptions $\Phi_t = \{\phi_m\}_{m=1}^{M_t}$ for every frame, where $M_t$ is the total number of interest points detected in the frame, and interest point $\phi_m$ is defined as $\phi_m = [\phi_m^x, \phi_m^y, \phi_m^d]$. $\phi_m^x$, $\phi_m^y$ denote its coordinates (i.e. those of the corresponding local region's centroid) on the image grid and $\phi_m^d$ is the local descriptor vector, e.g. an 128-element SIFT vector.

Temporal correspondence between an interest point $\phi_m \in \Phi_t$ and one interest point of the previous frame can be established by local search in a square spatial window of dimension $2 \cdot \sigma + 1$ of frame $I_{t-1}$, i.e. by examining if one or more $\phi_n \in \Phi_{t-1}$ exist that satisfy the following conditions:

$$|\phi_m^x - \phi_n^x| \leq \sigma \tag{1}$$

$$|\phi_m^y - \phi_n^y| \leq \sigma \tag{2}$$

$$d(\phi_m^d, \phi_n^d) \leq d_{sim} \tag{3}$$

where $\sigma$ is a constant whose value is chosen such that a reasonably-sized square spatial window is considered during local search, and d(.,.) is the Euclidean distance (which was also used in [1] for keypoint matching across different images). If multiple interest points satisfying Eqs. (1)-(3) exist, the one for which quantity $d(\phi_m^d, \phi_n^d)$ is minimized is retained. When such an interest point $\phi_n$ exists, the interest point $\phi_m \in \Phi_t$ is appended to the feature track where the former belongs, while otherwise (as well as when processing the first frame of the shot) the interest point $\phi_m$ is considered to be the first element of a new feature track.

Repeating the temporal correspondence evaluation for all pairs of consecutive frames in $S$ (and all interest points of the second frame of each such pair) results in the extraction of a set $\Psi$ of feature tracks, $\Psi = \{\psi_k\}_{k=1}^{K}$, where $\psi_k = [\psi_k^x, \psi_k^y, \psi_k^d]$. $\psi_k^d$ is the average descriptor vector of a feature track, estimated by element-wise averaging of all interest point descriptor vectors $\phi_m^d$ of the feature track as in [13], while $\psi_k^x$ is the corresponding time series of camera-motion-compensated interest point displacement in the x-axis between successive frames of $S$ where the feature track is present, and similarly $\psi_k^y$ for the y-axis. Thus, $\xi_k = [\psi_k^x, \psi_k^y]$ is the long-term trajectory of the interest point that generates the feature track: $\psi_k^x = [\psi_k^{x,t_1},\ \psi_k^{x,t_1+1},\ \ldots\ \psi_k^{x,t_2}]$ where $t_2 > t_1$ (and similarly for $\psi_k^y$). The values $\psi_k^{x,t}$ are estimated for any given $t$ by initially using the differences $\phi_m^x - \phi_n^x$, $\phi_m^y - \phi_n^y$ for all identified valid pairs of interest points between frames $I_t$, $I_{t-1}$ to form a sparse, non-regular motion field for the corresponding pair of frames; subsequently, the 8 parameters of the bilinear motion model, representing the camera motion, are estimated from this field using least-squares estimation and an iterative rejection scheme, as in [7]. Then $\psi_k^{x,t}$ and $\psi_k^{y,t}$ are eventually calculated as the differences between the initial displacement of the corresponding interest point's centroid between times $t-1$ and $t$, and the estimated camera motion at the location of the centroid.

The simple interest point matching between successive frames of $S$ that is used as part of the proposed feature track extraction process was chosen primarily for its simplicity; more elaborate techniques for tracking across frames can be used instead, if the added computational complexity is not a limiting factor.

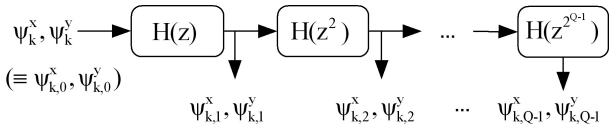


**Fig. 1**. Filter bank used for capturing motion at different time-scales.

### 3.2. LIFT representation

The extracted feature tracks are variable-length feature vectors, since the number of elements comprising $\psi_k^x$ and $\psi_k^y$ is proportional to the number of frames that the feature was successfully tracked in. This fact, together with other possible track artefacts (e.g. the extraction of partial tracks, due to failure in interest point matching between consecutive frames, occlusions etc.) make the matching of feature tracks non-trivial and render their current representation unsuitable for direct use in a Bag-of-Words approach. To this end, each motion trajectory is transformed to a fixed-length descriptor vector that attempts to capture the most important characteristics of the motion.

To capture motion at different time-scales, $\psi_k^x$ and $\psi_k^y$ are initially subject to low-pass filtering using a filter bank shown in Fig. 1, based on the lowpass Haar filter $H(z) = \frac{1}{2}(1 + z^{-1})$. This results in the generation of a family of trajectories, $\xi_{k,q} = [\psi_{k,q}^x, \psi_{k,q}^y]$, $q = 0, ..., Q-1$, as shown in Fig. 1, which due to the simplicity of the Haar filter are conveniently calculated as follows:

$$\psi_{k,q}^x = [\psi_{k,q}^{x,t_1+2^q-1},\ \psi_{k,q}^{x,t_1+2^q},\ \ldots\ \psi_{k,q}^{x,t_2}] \tag{4}$$

$$\psi_{k,q}^{x,t} = \frac{1}{2^q} \sum_{i=0}^{2^q-1} \psi_k^{x,t-i} \tag{5}$$

The $y$-axis elements of the trajectory are calculated similarly.

For any trajectory $\xi_{k,q}$, the histogram of motion directions at granularity level $\theta$ is defined as a histogram of $\frac{2\cdot\pi}{\theta}$ bins: $[0, \theta)$, $[\theta, 2\cdot\theta)$,..., $[2\cdot\pi-\theta, 2\cdot\pi)$. The value of each bin is defined as the number of elementary motions $[\psi_{k,q}^{x,t},\ \psi_{k,q}^{y,t}]$ of the trajectory that fall into it,

normalized by division with the overall number of such elementary motions that belong to the examined trajectory. $\lambda(\xi_{k,q}, \theta)$ is defined as the vector of all bin values for a given $\xi_{k,q}$ and a constant $\theta$.

Then, the initial trajectory $\xi_k$ can be represented across different time-scales as a fixed length vector $\mu_k$,

$$\mu_k = \Big[\lambda(\xi_{k,0}, \frac{\pi}{2}),\ \lambda(\xi_{k,1}, \frac{\pi}{2}),\ ... \lambda(\xi_{k,Q-1}, \frac{\pi}{2})$$
$$\lambda(\xi_{k,0}, \frac{\pi}{4}),\ \lambda(\xi_{k,1}, \frac{\pi}{4}),\ ... \lambda(\xi_{k,Q-1}, \frac{\pi}{4}) ...$$
$$\lambda(\xi_{k,0}, \frac{\pi}{2^J}),\ \lambda(\xi_{k,1}, \frac{\pi}{2^J}),\ ... \lambda(\xi_{k,Q-1}, \frac{\pi}{2^J})\Big] \quad (6)$$

where $J$ is the number of granularity levels. The corresponding Local Invariant Feature Track (LIFT) is defined as

$$LIFT(\psi_k) = [\psi_k^d,\ \mu_k] \quad (7)$$

The LIFTs of a video shot can be used for generating a Bag-of-Words model that will essentially describe the shot in terms of classes of "similarly-moving, visually-similar local regions", rather that simply "visually-similar local regions" (detected by either spatial or spatio-temporal interest point detectors), as in the current state-of-the-art, e.g. [8, 11].

### 3.3. Invariance concerns

The definition of the LIFT representation was guided by the need to introduce to the extent possible some invariance with respect to the scale and direction of the extracted tracks. Starting with the interest point detection and description in the 2D, the SIFT method was used, due to its well-documented [1, 3] desirable invariance properties; other similar methods [4, 5] could also be used instead. Concerning the feature track extraction, camera-motion-compensated trajectories were estimated and employed to ensure that the final LIFT representation will not be affected by camera motion.

In the subsequent representation of the tracks by histograms, only the direction of each elementary motion of the track was employed, rather than the direction and magnitude of it. This was done for introducing some degree of invariance to image scale, since the same motion (e.g. a person picking up the phone) will result in different motion vector magnitudes depending on the focal length of the camera and its distance from the plane of the motion; on the contrary, the direction of motion is not affected by these parameters.

Histograms at various time-scales were selected for representing the tracks, instead of e.g. comparing the overall displacement of the interest point along the track, to allow for partial matches when considering partial tracks (i.e. when the beginning and end of the different extracted tracks that correspond to the same class of actions do not coincide with each other and with the actual beginning and end of the depicted action). Although the adopted solution may be non-optimal, the reliable matching of partial tracks would otherwise require the use of a computationally expensive optimization-based technique for evaluating the similarity of them, in place of the Euclidean distance typically used in K-means when creating the "words" used in the Bag-of-Words approach.

The use of motion direction histograms at different granularity levels $\theta$ (instead of using a single histogram with a high number of bins) aims at allowing again for partial matches between tracks using a simple metric (i.e. L1/L2 rather than e.g. the Earth Mover's Distance), in the case of small variations in the direction of motion. When considering only a very fine granularity level $\theta$, significant such variations between similar shots could be caused by even small differences in camera angle/viewpoint. The combined use of multiple (from coarse to fine) granularity levels can alleviate this effect to some degree. Alternatively, the weighted assignment of every elementary motion to more than one neighboring bins, when constructing each motion direction histogram, could be employed.

## 4. EXPERIMENTAL RESULTS

In the experimental evaluation of the proposed LIFT features, the fully annotated TRECVID 2007 training and test datasets were employed, comprising 50 and 50 hours of video, and 18120 and 18142 shots, respectively. The 20 high-level features that were defined on this dataset for the TRECVID 2009 contest were used for the preliminary evaluation of the proposed approach.

In extracting the proposed LIFT representation of the video shots, the temporal sub-sampling parameter $a$ was set equal to 3. This represents a good compromise between the need for accurately establishing the SIFT point correspondences from frame to frame (which calls for a low value of $a$, ideally 1) and the need for speeding up the feature extraction process. For each frame of the temporally subsampled sequence, the method of [1] was used for interest point detection and the description, resulting in 128-element vectors for the local region of each interest point. Parameter $\sigma$ defining the local window where correspondences between SIFT descriptors are evaluated was set to 20, and parameter $d_{sim}$ used for evaluating the similarity of SIFT descriptors in different frames was set to 40000. Four different timescales ($Q = 4$) and three granularity levels (i.e. $J = 3$ in Eq. (6)) were used for representing the trajectory information of the extracted feature tracks. As a result, a 240-element vector was used for representing each LIFT feature.

For comparing LIFT with key-frame based SIFT, the median frame of each shot was selected as a key-frame and SIFT descriptors were extracted, as above, for each key-frame. For both SIFT and LIFT, the popular Bag-of-Words (BoW) model was used for estimating a single descriptor vector for each shot; in our experiments, the number of visual words was set to 500. Support Vector Machine (SVM) classifiers producing a fuzzy class membership degree in the range [0,1] were then used for evaluating the relevance of each shot of the TRECVID 2007 test dataset with every one of the considered high-level features, exploiting two different combinations of BoW models: i) SIFT-based BoW, and ii) concatenation of the aforementioned SIFT-based BoW with a LIFT-based BoW, the latter being constructed using the LIFT descriptors of section 3.2. In any case, the corresponding SVM classifiers were previously trained using the TRECVID 2007 training dataset and the common annotation; for each combination of BoW models and each high-level feature, a single SVM was trained independently of all others.

High-level feature extraction results (average precision@2000 [6]) for both SIFT-based BoW features and features resulting from the concatenation of SIFT- and LIFT-based BoW are shown in Fig. 2. It can be seen that introducing the proposed LIFT representation results in significantly higher precision than using solely the SIFT one, particularly when considering high-level features that have a strong temporal dimension (e.g. "people-dancing", "person-playing-soccer", etc.). Overall, the SIFT-based BoW resulted in a mean average precision (MAP) of 0.0538, whereas the combination of SIFT- and LIFT-based BoW in a MAP of 0.0712, representing an increase of the former by approximately 32%. Considering only the 6 high-level features of Fig. 2 that have a strong temporal dimension (i.e. features 5, 6, 7, 9, 11, 13), the use of the proposed combination of SIFT- and LIFT-based BoW leads to an increase of MAP by 63.8% over using the SIFT-based BoW alone.

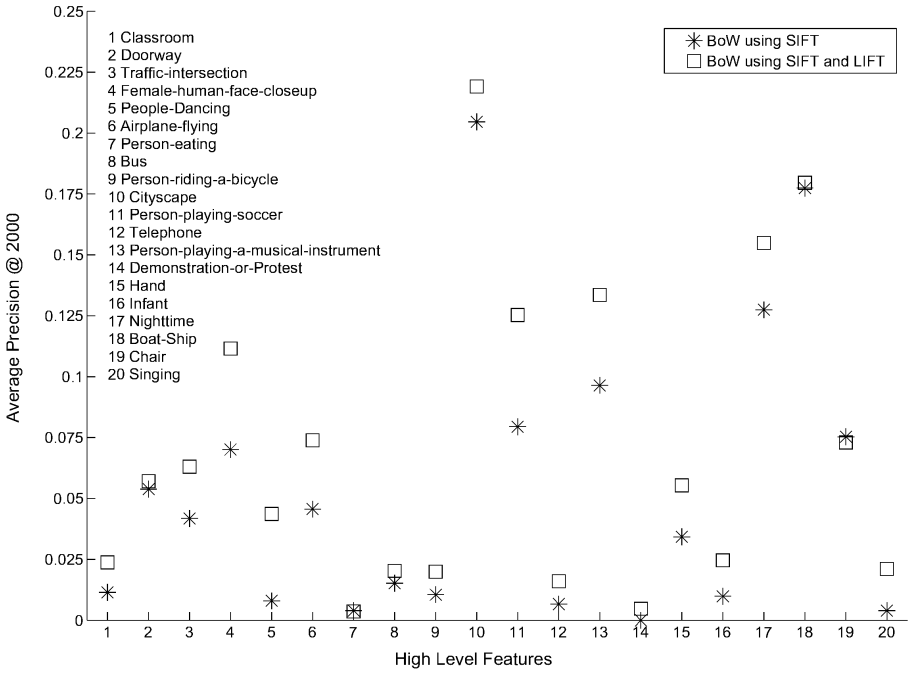


**Fig. 2**. High-level feature extraction results on the TRECVID 2007 dataset.

## 5. CONCLUSIONS

In this work the extraction and representation of Local Invariant Feature Tracks was proposed for jointly capturing the spatial attributes and the long-term motion of local regions in video. In combination with the BoW technique, the proposed LIFT representation can be used for generating Bags-of-Spatiotemporal-Words models, thus introducing information about the temporal evolution of each video shot in BoW. Preliminary experimental evaluation of the proposed approach on the corpus of TRECVID 2007 revealed its potential for high-level feature extraction from video.

## 6. REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Computer Vision*, vol. 60, pp. 91–110, 2004.

[2] C. Dance, J. Willamowski, L.X. Fan, C. Bray, and G. Csurka, "Visual categorization with bags of keypoints," in *Proc. ECCV Int. Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic, May 2004.

[3] K. Mikolajczyk and C. Schmid, "Performance Evaluation of Local Descriptors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 16151630, 2005.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[5] G. J. Burghouts and J.-M. Geusebroek, "Performance Evaluation of Local Colour Invariants," *Computer Vision and Image Understanding*, vol. 113, pp. 48–62, 2009.

[6] A. F. Smeaton, P. Over, and W. Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements," in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran, Ed., pp. 151–174. Springer, 2009.

[7] V. Mezaris, I. Kompatsiaris, N.V. Boulgouris, and M.G. Strintzis, "Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 606–621, May 2004.

[8] C.G.M. Snoek, K.E.A. van de Sande, O. de Rooij, and et.al., "The MediaMill TRECVID 2008 Semantic Video Search Engine," in *Proc. TRECVID 2008 Workshop*, USA, Nov. 2008.

[9] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, "Video Event Classification using String Kernels," *Multimedia Tools and Applications*, 2010.

[10] I. Laptev, "On Space-Time Interest Points," *Int. J. of Computer Vision*, vol. 64, no. 2/3, pp. 107–123, 2005.

[11] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Int. J. of Computer Vision*, vol. 79, no. 3, pp. 299–318, Sept. 2008.

[12] H. Zhou, Y. Yuan, and C. Shi, "Object tracking using SIFT features and mean shift," *Computer Vision and Image Understanding*, vol. 113, pp. 345–352, 2009.

[13] A. Anjulan and N. Canagarajah, "A Unified Framework for Object Retrieval and Mining," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 19, no. 1, pp. 63–76, Jan. 2009.

[14] N. Moenne-Loccoz, E. Bruno, and S. Marchand-Maillet, "Local Feature Trajectories for Efficient Event-Based Indexing of Video Sequences," in *Proc. CIVR*, Tempe, USA, July 2006.